THE SURVEY OF INCOME AND PROGRAM PARTICIPATION

AN EXPLORATION OF THE APPLICABILITY OF HAZARDS MODELS IN ANALYZING THE SURVEY OF INCOME AND PROGRAM PARTICIPATION:

LABOR FORCE TRANSITIONS

No. 254

Kathleen S. Short
Karen A. Woodrow
U.S. Census Bureau

U.S. Department of Commerce U.S. CENSUS BUREAU

DRAFT: 7/29/85

An Exploration of the Applicability of Hazards Models in Analyzing the Survey of Income and Program Participation: Labor Force Transitions

by

Kathleen s.Short and Karen A. Woodrow

Population Division

Bureau of the Census Washington, D.C.

Paper to be presented at the Joint Statistical Meetings,

Las Vegas·, Nevada, August 1985.

An Exploration of the Applicability of Hazards Models in Analyzing the Survey of Income and Program Participation: Labor Force Transitions

by

Kathleen S. Short and Karen A. Woodrow

The Survey of Income and Program Participation is designed to provide a dynamic, longitudinal view of educational, labor force, and demographic activity both during the two and one-half years of data collection and for individuals' lifetime experience prior to the initial data gathering point. This paper is the first in a series which will explore the applicability of hazards modeling techniques for the analysis of demographic, sociological and economic transitions using SIPP data.

Hazards models examine the time paths of events that occur to individuals. These models, also known as multivariate life tables or life tables with covariates, analyze the effects of independent variables upon the time dependent risk of experiencing an event, e.g.; death, divorce, marriage, or childbirth. Hazards models present a view of the factors related to both the number and the timing of such transitions of interest.

Information collected in the SIPP lends itself to studies of many types of transitions. Changes in family composition, marital status, program participation, labor force participation, and other such life events, are substantially documented over the sample period. Also, detailed persona1 hist1lries for individua1s covering 1abor force experiences, program participation and household composition changes are collected for the same individuals.

For the purpose of this study the event of interest is the transition from the state of unemployment to employment. Differences in probabilities of moving from one state to another, and the timing of such movements by characteristics of individuals will be examined. Covariates inthe model need not be limited to characteristics of an individual, which areunchangingover time, such assex and race, but may alsoinclude characteristics measured at the beginning of the study and measures of change in these characteristics as ascertained over time. Hazards modeling techniques are also ideal for dealing statistically with problems of right-censored data, that is, the problems caused by the varying time at risk for individuals in the study due to sample attrition or the ending of the study.

At the present stage of data completion, information is 1imited to a four month period, with weekly data on labor force status in particular. Using this information it is possible to pinpoint the week in which an individual experienced a transition from one state to the other.

This study explores the adequacy of the SIPP for this type of analysis and suggests ways in which it could be made more suitable. This project also lays the groundwork for future application as more information becomes available. The work history fixed topical module, for example, provides a useful event history suitable to these purposes. Beyond the need for information about duration of the state in which an individual enters the sample, it is also desirable to know previous histories of event occurrences. Knowledge of the number of 1ike events experienced by each individual is important for the employment of hazard rate models.

I. INTRODUCTION

This study uses Wave 1 of the Survey of Income and Program Participation (SIPP) to illustrate the application of hazards modeling techniques. Wave 1 of the survey covers four months of 1abor force, income, and program participation information forapproximately 53,000individuals in four different rotation groups from June 1983 to December 1983.

The intent of the study is primarily to serve as a hands on "experiment" in the usability of these data in this type of analysis, for the purposes of identifying deficiencies and shortcomings of the survey design and making recommendations for future changes that would facilitate this and other types ofdynamic analysis.

The analysis which is described in the fo11owing paper is meant to serve only as an illustration. For this reason, a formal theoretical discussion is deemed inappropriate since no formal hypotheses are intended to be tested. Also, no inferences about the population are meant to be drawn from the results of the analysis. The analysis focuses on the labor force transitions primari1y due to the availability of the data, and the empirical specification is primarily ad hoc. Even so, labor economists and others familiar with issues important to unemployment theory will appreciate the richness of this set of data for this type of research.

In a sense, the data described herein represent the SIPP in its most primitive state. As more waves of information such as work history, marital history, education history and the event history modules become available, application of hazards models will be enhanced. Future changes in the questionnaire design proposed as a result of this study, and others. wi11 further improve the usability of SIPP for dynamic ana1ysis.

II. DESCRIPTION OF METHODS

Hazards models examine the time paths ofevents that occur to individuals. Thesemodels,also known as multivariate life tables with covariates, analyze the effects of independent variables upon the time dependent risk of experiencing an event.

This method of analysis is useful because it incorporates the amount of time spent in a particular state in parameter estimation. More traditional methods do not do this. Capturingthe effect of duration ina status on the probability of leaving that status is made possible by the kind of data available in a longitudinal survey such as SIPP.

We begin with a population of individuals for each of whom was observed either the time to failure, time to loss, or the end of the sample period. (1) In this application "failure" is synonymous with employment, "loss" would be non-reporting of employment status in the 2nd, 3rd, or 4th months (this does not occur in our sample). We assume T, the time to failure, to be a random variable with values [0, T*], probability density function f(t) on [0, T*] such that,

$$(1) \qquad f(t) \geq 0 \forall t$$

$$(2) \qquad \text{and} \int_{-\infty}^{\infty} f(t)dt = 1$$

where f(t) is approximately the probability of failure between time t and t+dt.

We then define a function G (t) and a family of conditional probabilities such that G (t) = 1 - F(t), called the survivor function, and

$$(3) \qquad G(t) = 1 - F(t) = 1 - s\varpi f(\tau)d\tau$$
$$= \int_{t}^{T*} f(\tau)d(\tau) \quad for \quad 0 \le t \le T*$$

is the probability of still being in a status at time t. From this we can calculate the hazard function, [need symbol from file here](t), which is the probability of failing at time t given survival to that point. Let

$$(4) \qquad \lambda(t) = (f(t)dt)/G(t) \text{ and } t \le t < t + dt \le T*$$

The objective of the analysis is to use microdata to estimate hazard rates, i.e. to estimate the probability of becoming employed at time t given that an individual has been unemployed unti1 time t. The hazard rate itself is unobservable, however, for each individual we observe characteristics Z_1,…., Z_p and we can characterize the relationship between the distribution of failure time and the vector Z.

In this study the event of the interest is "becoming employed", or “getting a job”. Individuals who are “at risk of becoming employed" are examined and parameters affecting the time to failure (becoming employed) are estimated.

Information about employment status in SIPP is reported in discrete intervals, in this , case, weekly intervals. For this reason, a discrete time method for estimating hazard rates is employed.

Following A11ison (.2), we have n individuals, and time intervals such that t=l at the beginning of observation and t=t; when the event of question occurs or the spell is censored. Assume T is a discrete random variable and Pit= PR[Ti= t[Tiλt, Zit] is a discrete time hazard rate.

Allison has shown that the log likelihood function for the hazard rate of person i at time t is

$$L = \sum_{i=1}^{n} \sum_{j=1}^{ti} Y_{ij} \, log[\frac{P_{ij}}{1-P_{ij}}] + \sum_{i=1}^{n} \sum_{j=1}^{ti} log(1 - P_{ij})$$

where $Y_{it} = 1$ uncensored; 0 censored

This log likelihood function is similar to that for a regression of a dichotomous dependent variable which takes the value of 1 if person I experiences the event at time t, and 0 otherwise. Note, however, that the summation is carried out over both persons and time periods. This requires restructuring of the data such that there is an observation for each person in each discrete time period until the next event occurs. The appropriate unit of analysis for this study is person weeks rather than persons.

Thus, we use the logistic functional form for Pit

$$(5) \qquad P_{it} = 1/[1 + exp(-\alpha_t - \beta'_t Z_{it})]$$

which may be written

$$(6) \qquad log(\frac{P_{it}}{1-P_{it}}) = \alpha_t + \beta'_t Z_{it}$$

The observations used to estimate the parameters of this hazard function are person weeks spent in job search. Maximum 1ike1ihood estimates of the parameters are obtained using a 1ogit regression. This procedure yields estimates of Bt from which inferences about the importance of the characteristic, Zi to the hazard rate, Pit, can be made.

III. DESCRIPTION OF DATA

Information collected in the SIPP lends itself to studies of many types of transitions. Data for this study of transitions in employment status are drawn from the labor force and recipiency section of the core questionnaire. In this section individuals respond that they worked in all weeks of the 4-month period covered in the first wave of the survey, or they point out, using flashcards, which weeks of the period they were or were not employed. Using this information it was possible to pinpoint the week in which a transition from one state to the other was experienced. Using these data, probabilities of experiencing transitions by characteristics such as marital status, region, etc. are estimated.

To do this, durations in status must be constructed. Information about the time at which the spell of unemployment began is important. Without this information the observation is

1eft censored. Left censored observations bias the estimated probability of the event occurring because left-censorship is not independent of the occurrence of the event. For this study left-censored observations are deleted by selecting only individuals for whom the beginning of an unemployment spell has been observed in the sample period. This process does introduce considerable selection bias into the estimator.

For this study, only the first spell of unemployment encountered in the sample period is studied. The sample consists of males aged 16 to 65 years who were employed in the first week of the sample period and who subsequently experienced a period without a job during which time they looked for work. Individuals who were in the armed forces, self-employed or who were college students were deleted from the sample, leaving a total of 272 individuals.

Among these individuals, responses to the question "In which of these weeks were you looking for work or on 1ayoff from a job?" were mixed. A number of individuals reported spending time away from a job both looking and not looking for a job. For example, a respondent might report not looking for a job in the week immediately following the loss of job, looking for a job for a week or two subsequently and not looking again for a week before employment resumed. These individuals are classified here as “unemployed" and the entire duration is counted as a spell of unemployment.

This type of inconsistent response has been noted and discussed in past research on unemployment. Clark and Summers (1979) suggest that "looking for work" is an ambiguous concept. They discuss implications from the 1961-19&6 CPS reinterview program that,

"many of those not in the labor force are in situations effectively equivalent to the unemployed."

Clark and Summers documented changes in and out of unemployment and not in labor status by month from the 1976 CPS over the months May, June, July and August. They state that it is implausible that individuals looking for work in May, not looking in June, and again "looking in July have significantly changed their job seeking intentions over the period, and that these responses ", … reflect the ambiguity and arbitrariness inherent in any definition of labor force activity."

We suggest that the conclusions of Clark and Summers are even more applicable to these data which reflect weekly changes from unemployment to out of labor force. It seems even more implausible that an individual looking for work in weeks 1, not looking in weeks 2 and 3, and looking again in week 4, has changed his or her labor force intentions. Thus, these individuals were treated as "unemployed" across the duration of being without a job.

Calculation of duration of the first observed spell of "unemployment" was done by summing across weeks marked as "without a job" and "looking for work," even if job search was intermittent, until a week with a job was encountered or until the end of the survey period being examined. These calculated durations are shown along with the number of cases with one completed spell in Table 1.

Table 1: Durations of First Observed Spell of Unemployment		
Duration in Weeks	Completed	Right Censored
1	17	8
2	26	15
3	22	10
4	11	18
5	10	10
6	8	13
7	7	13
8	2	7
9	2	10
10	5	6
11	3	6
12	1	6
13	2	15
14	0	5
15	0	8
16	0	6
Total	116	156

IV. RESULTS

The fol1owing tables present the results from the regression analyses. Several independent variables are included throughout.

Age is entered as a categorical variable. Three age categories were constructed; less than 26 years (Al), over 54 years (A3), and the omitted category, 26 to 54 years of age. Race is included, the included dummy variable representing white; black and others are the omitted category.

Receipt of three types of unemployment compensation are included asone categorical variable, representing state unemployment compensation, supplemental unemployment benefits, or other unemployment compensation.

The excluded category represents non-receipt of any of these types of unemployment compensation.

Education is represented as three categorical variab1es. Included covariates are those who have not completed high school and those with college education. The excluded category is for individuals with a high school education only. Total earned income by this person in the first month is used to calculate relative importance of this person's job to household income. The ratio of this person's earned income to total household income in month one of the observed period is included as a measure of the contribution of this person’s job to the pool of resources available as a measure of the contribution of this person's job to the pool of resources available to the household in which he resides.

Receipt of foodstamps or AFDC payments are included as one categorical variable representing receipt of transfer income. These, as other types of transfer incomes received, have been found in past research to have increased the length of unemp1oyment spe11s, in so far as they allow the individual to search longer for a more agreeable job offer.

First we show the results of estimating a model in which time is merely a control variable and the assumption is made that time is related linearly to the dependent variable.

Table 2 presents results of an ordinary least squares regression estimating effects of age, race, education, incomes received and total duration of unemployment on whether or not a person in this sample becomes employed within the sample period. The dependent variable is a dichotomous variable that takes the value of one if a person became reemployed by the interview date. This estimation used only the 272 person unstructured data set. Other than the inappropriate use of ordinary least squares this estimation procedure has some very obvious problems. The estimate coefficients suggest that a college education affects reemployment. None of the independent variables are important except duration unemployed. The estimated coefficient for duration suggest a slightly negative relationship between duration unemployed and reemployment, i.e. the 1onger a sample person is unemployed, the less likely that person is to be reemployed. However, it is also true, due to right censored observations, that the longer a person is unemployed, the less likely we are to observe reemployment within our 16 week window of observation. Thus, the significant relationship expressed between duration in status and the occurrence of the event is rea11y only an artifact of the censoring in the data.

In order to estimate the discrete hazard rates, the data set is restructured in order that the unit of analysis is person weeks rather than persons.

For example, an individual aged 24, white, with a high school education is observed to have become unemployed in the second week of the sample period, remain unemployed for 4 weeks, and become reemployed in the 7th week of period. The observation for this person is shown in Panel A of Table 3.

In the data transformation process this individual would become 4 person weeks of unemployment as shown in Panel B of Table 3. In addition dummy variables are constructed representing each of the 16 weeks of the observation period. These categorical variables take the value of 1 if the person was unemployed for at least that duration and 0 otherwise. Another categor1cal variable is calculated for each person week which takes the value of 1 if the individual were employed in this person week and 0 if not.

Table 2: Ordinary Least Squares Regression n= 272 Persons		
DEPENDENT VARIABLE	1 Becomes Reemployed	
	0 Otherwise	
	β	**β/σ**
AGE		
16-24 years	-0.02696	0.423
55-65 years	0.06728	1.249
RACE		
White	0.10715	1.311
EDUCATION		
No High School	0.81992	1.237
College	0.13878	2.000
TRANSFER INCOME	0.14414	1.223
UNEMPLOYMENT BENEFITS	0.06232	0.932
BREADWINNER	0.06502	0.692
DURATION UNEMPLOYED	-0.04658	6.762
CONSTANT	0.57031	1.221
R_2	0.19448	

Table 3: Translating Data for Sample Person to Person Week Observations						
Panel A		**Panel B**				
Variable	Value	Variable	Person Weeks			
AGE			1	2	3	4
16-24 years	1	AGE				
55-65 years	0	16-24 years	1	1	1	1
RACE		55-65 years	0	0	0	0
White	1	RACE				
EDUCATION		White	1	1	1	1
No High School	0	EDUCATION				
College	0	No High School	0	0	0	0
TRANSFER INCOME	0	College	0	0	0	0
UNEMPLOYMENT BENEFITS	0	TRANSFER INCOME	0	0	0	0
BREADWINNER	0.42	UNEMPLOYMENT BENEFITS	0	0	0	0
DURATION	4	BREADWINNER	0.42	0.42	0.42	0.42
REEMPLOYED	1	t_1	1	0	0	0
		t_2	0	1	0	0
		t_3	0	0	1	0
		t_4	0	0	0	1
		t_5	0	0	0	0
		t_6	0	0	0	0
		t_7	0	0	0	0
		t_8	0	0	0	0
		t_9	0	0	0	0
		t_{10}	0	0	0	0
		t_{11}	0	0	0	0
		t_{12}	0	0	0	0
		t_{13}	0	0	0	0
		t_{14}	0	0	0	0
		t_{15}	0	0	0	0
		t_{16}	0	0	0	0

The means and standard deviations in Table 4 are based upon the data set of person weeks observed as unemployed at the beginning of each week and either reemp1oyed or observed for a subsequent week in unemployed status.

The results of the logistic regression are presented in Table 5.

It wi11 be noted here that the maximization process used to estimate the model which includes all sixteen weeks of information did not converge. For this reason all person weeks beyond week 13 were selected out of the data set, and the time variab1es, accordingly, were omitted from the model. The logistic regressions were estimated using 1626 person weeks of observation.

In model A the probability of becoming employed is expressed as a function of age, race, level of education, transfer incomes received, as well as proportion of total income this person contributed in his last job. Estimated coefficients show that the probability of employment is significantly decreased for persons in the sample, if a person is young or receiving unemp1oyment benefits, whereas, a person is more likely to leave the state of unemployment if they are white or have a college education. Receiving food stamps or AFDC does not affect the probability of becoming reemployed.

Model B includes time variables as covariates in the estimated equation, expressing the probability of becoming emp1oyed as a function of the characteristics in Model A and duration of the unemployment period. The estimated coefficients on the time variables, significant and positive for 5 weeks of unemployment, suggest that the probability of becoming employed rose in the fifth week of the unemployment period, after which/time in

status has no effect. The increase in the ca1cu1ated chi-square statistic was insufficient to conclude that the time variables added significantly to the exp1anatory power of the model.

It appeared that the time variables representing longer durations in status may not be significant because of the high degree of right censoring present in these data. Allison (5) suggests testing for sensitivity to violations of the assumption of independence between censoring and the occurrence of the event by assigning values of 1 to the dependent variable in the last week of observed unemployment, whether the observation is censored or not. This has been done in Model C. Note that in this estimation nearly all the estimated coefficients for the time variables are significant and positive, suggesting that the large number of right censored observations seriously affects the estimation of coefficients of the higher level time variables. Slight changes in the estimated parameters for other covariates suggest that, for this analysis, right censoring is not independent of the hazard of becoming reemployed. It is suggested that a longer period of time in sample would alleviate this sensitivity to right censoring.

Table 4: Means and Standard Deviations for Independent Variables and Dependent Variable		
VARIABLE	MEAN	STANDARD DEV.
Dependent Variable	0.076	0.265
AGE		
15-25 years	0.564	0.496
55-65 years	0.023	0.149
RACE		
White	0.831	0.375
EDUCATION		
No High School	0.299	0.458
College	0.244	0.430
BREADWINNER	0.378	0.485
DURATION UNEMPLOYED		
2 weeks	0.148	0.356
3 weeks	0.124	0.329
4 weeks	0.105	0.306
5 weeks	0.087	0.282
6 weeks	0.075	0.264
7 weeks	0.062	0.242
8 weeks	0.050	0.219
9 weeks	0.045	0.207
10 weeks	0.038	0.191
11 weeks	0.031	0.174
12 weeks	0.026	0.159
13 weeks	0.022	0.145
14 weeks	0.011	0.106
15 weeks	0.008	0.091
16 weeks	0.004	0.060
n= 1665 persons		

Table 5: Logistic Regression Results						
Dependent Variable: LOG ODDS OF BECOMING REEMPLOYED	MODEL A		MODEL B		MODEL C	
	COEFFICIENT	COEFFICIENT/ S.E.	COEFFICIENT	COEFFICIENT/ S.E.	COEFFICIENT	COEFFICIENT/ S.E.
Age						
16-25 years	-0.2413*	-2.29	-.2399*	-2.26	-0.1535*	-1.92
55-65 years	0.2799	0.84	0.3174	0.95	0.0609	0.23
Race						
White	0.3355*	2.12	.3321*	2.09	0.1208	1.19
Education						
No High School	0.2258*	1.96	0.2083	1.79	0.0827	0.98
College	0.2772	2.35	0.2668	2.25	0.0727	0.82
Transfer	-0.0208	-0.12	-0.0259	-0.15	-0.2687*	-1.80
Unemployment Benefits	-0.2791	-2.40	-0.2662*	-2.21	-0.3747*	-4.13
Breadwinner	0.2447	2.35	0.2192*	2.06	0.3442*	4.12
Duration Unemployed						
2 weeks			0.2998	1.83	0.3547	2.59
3 weeks			0.3196	1.88	0.3332	2.31
4 weeks			0.0443	0.22	0.3949	2.66
5 weeks			0.5011	2.84	0.3016	1.86
6 weeks			0.0847	0.38	0.4323	2.67
7 weeks			0.1487	0.63	0.5511	3.31
8 weeks			-0.3891	-1.02	0.2144	1.03
9 weeks			-0.3118	-0.82	0.4595	2.38
10 weeks			0.2661	0.99	0.5256	2.62
11 weeks			0.1063	0.32	0.5351	2.48
12 weeks			-0.3771	-0.72	0.4908	2.08
13 weeks			0.0663	0.17	1.2636	6.23
Constant	3.425	18.52	3.272	14.78	3.7150	23.42
Chi-Square	1657.5		1673.3		1647.6	
Degrees of Freedom	1627		1605		1605	

DISCUSSION

This analysis, as was stated in the introduction, was intended primarily to serve as an illustration of a particular type of analysis using SIPP Wave l data. Much more fruitful results would have been obtained if the sample period had covered a longer period of time, allowing observation of both more individuals experiencing spells of unemployment and more completed spells. The data available from the first wave only were sparse in person weeks of unemployment.

This was exacerbated by the left censoring inherent in the survey design. It has been shown by Heckman and Singer (3) that only the distribution of duration of spells initiated after the beginning date of the sample is invariant to the sampling plan. They suggest, if there are no unobserved variables, that the analysis be confined to spells which begin during the period of observation. Thus, individuals in a state of unemployment in the first week of the sample period were excluded from the sample. However, limiting the sample to those for whom the window of observations recorded the beginning of the spell of unemployment has not only severely limited the degrees of freedom, but has also introduced sample selection bias.

Restricting the sample to individuals who were employed in the first week of the sample period systematically eliminated those for whom unemployment spells are of long duration. Individuals unemployed for a long period of time had a lower probability of being found to be employed in the first week of the sample period.

As a result of this type of study and input from other researchers, changes in the design of the survey questionnaire have been proposed for the 1986 panel. The changes focus on reducing the amount of left--censoring in the data. This will be done by

adding a personal history module to the second interview covering a wide variety of subject areas. These include marital, fertility, education, employment and recipiency histories. All include questions about the beginning of statuses recorded in the initial interview.

CONCLUSIONS

Conclusions that may be drawn from the above exercise in the application of hazards modeling to SIPP employment status information must include an appreciation for the amazing richness of these data. The amount of data available about individuals, their households, level and types of income, availability of resources, and personal characteristics is impressive, even in only the first wave of data collection. We have been able to apply a dynamic method of social research, the possibilities of which will be greatly enhanced as the survey is continued and improved.

Many of the problems encountered in this study will be alleviated by the new addition of information already collected and available. Further problems, such as left censored data, are being addressed by proposed changes in questionnaire.

The relatively simple application of the discrete time hazards model presented here serves to illustrate and point the way t-0 the more theoretically rigorous and analytically sophisticated kinds of research that will be made possible with the SIPP.

BIBLIOGRAPHY

1. Cox, D.R. Journal of the Royal Statistical Society, Series B, Vol. 34, No. 2 (1972) "Regression Models and Life-Tables"
2. Allison, Paul D. Sociological Methodology (1982) "Discrete Time methods for the Analysis of Event Histories."
3. Clark, Kim B. and Summers, Lawrence H, Brookings Paper on Economic Activity, 1:1979. "Labor Market Dynamics and Unemployment: A Reconsideration".
4. Heckman, James J. and Singer, Burton. Journal of Econometrics 24 (1984) "Econometric Duration Analysis."
5. Allison Paul D. "Event History Analysis: Regression for Longitudinal Event Data". Sage University Papers, Series/Number 07-046, Sage Publications, Beverly Hills and London, 1984.